ISSN 1561-5359. Штучний інтелект, 2017, № 2

УДК 004.93

А.О. Щерба
Київський національний університет імені Тараса Шевченка, Україна
вул. Володимирська, 60, м. Київ, 01033

ОПТИМІЗАЦІЯ НАБОРІВ ДАНИХ ПРИ ОЦІНЮВАННІ ВАРТОСТІ ОБ'ЄКТІВ РИНКУ НЕРУХОМОСТІ

A.O. Shcherba
Taras Shevchenko National University of Kyiv, Ukraine
60, Volodymyrska str., Kyiv, 01033

DATASETS OPTIMIZATION FOR REAL ESTATE VALUATION

У статті розглянуто питання підвищення якості наборів даних, що збираються з відкритих джерел. Проаналізовано головні особливості таких наборів даних, запропоновано шляхи їх вирішення. Ці підходи було протестовано на різних алгоритмах для оцінювання вартості нерухомості і серед них було визначено найкращі.

Ключові слова:, оцінка вартості нерухомості, навчальна вибірка, штучні нейронні мережі

The article deals with the problem of improving quality of real estate datasets which are created from the open sources of information. Main problems of such datasets are analyzed and some ways to solve them are suggested in this article. These solutions were tested on different algorithms for real estate valuation and the best ones were selected.

Keywords: real estate valuation, training dataset, artificial neural networks.

Вступ

Задача оцінювання вартості об'єктів нерухомості є дуже чутливою до розташування цих об'єктів [1], оскільки в різних країнах і навіть регіонах ціноутворюючими є зовсім різні фактори і їхній вплив так само сильно відрізняється залежно від регіону або країни.

Наприклад, в Україні площа квартири є ключовим фактором при оцінюванні вартості квартир, фактично всі інші характеристики формують ціну за квадратний метр, а вартість отримується як її добуток на площу квартири. Тобто, залежність між вартістю і площею є лінійною. У тому ж дослідженні [2], у якому порівнювалися різні підходи для даних зі штату Кентуккі (США), площа об'єкта перетворювалася на порядкове число (від 1 до 4), тобто залежність все ще присутня, проте значно менша.

Така різниця у впливі ціноутворюючих факторів на остаточну вартість призводить до того, що модель, яка навчалась на даних, з одного місця неможливо використовувати для оцінювання об'єктів у інших місцях. Але ця задача, в першу чергу, практична, тому розроблена система повинна оцінювати об'єкти на певній наперед заданій території [3],[4].

Постановка проблеми

Через це постає проблема отримання навчальних даних, що притаманна усім алгоритмам оцінювання об'єктів нерухомості. В Україні ця проблема погіршується тим фактом, що у нас в країні відсутній єдиний централізований відкритий реєстр з інформацією щодо купівлі-продажу об'єктів нерухомості. Тобто, немає жодних способів отримати реальні характеристики і вартість нерухомості.

Єдиним відкритим джерелом інформації щодо цих даних є об'яви з продажу нерухомості. Проте, вони мають деякі недоліки:

- заявлена вартість об'єктів не є остаточною і може бути знижена під час фінальних торгів;

- деякі з оголошень не відповідають заданим шаблонам і можуть не мати важливих характеристик, наприклад, площу або кількість кімнат;
- деякі з характеристик є дуже суб'єктивними і не можуть бути використані для об'єктивного оцінювання цього об'єкта, наприклад, стан квартири, вид з вікна і т. ін.;
- в об'яві можуть бути помилки, наприклад, неправильно вказана валюта, об'яви з оренди нерухомості знаходяться у розділі з продажу нерухомості.

Серед переваг цього джерела можна виділити наступне:

- за наявності певного формату повідомлення можна автоматично виділяти з нього основні характеристики і вартість;
- їхня відкритість означає, що інформацію з них можна збирати автоматично без додаткового опрацювання людиною;
- якщо джерело є досить популярним, то в ньому міститься багато оголошень і їхня кількість постійно зростає, що дозволяє отримати великий набір навчальних і тестових даних;
- у нових об'явах міститься актуальна інформація, а тому можна завжди підтримувати актуальність зібраних даних, видаляючи старі елементи і додаючи нові.

Як ми бачимо, за наявності джерела з об'явою, що відповідає поставленим вимогам (має велику кількість об'яв, що відповідають заданому формату), можна автоматизувати процес створення навчальної вибірки даних.

У такому випадку залишається лише розібратися з недоліками, оскільки вони є досить суттєвими.

Перший недолік неможливо вирішити об'єктивно, оскільки реальні суми угод невідомі, проте, по-перше, отримання навіть цієї вартості пропозиції є вже непоганим результатом, а, по-друге, цю різницю між ринковою вартістю і вартістю об'яв можна зменшити за допомогою статистичних методів, оскільки, зазвичай, ця величина у відсотках є достатньо фіксованою, тобто, зменшивши початкову вартість, можна отримати незміщену оцінку.

Другий недолік можливо вирішити лише відкиданням об'єктів, що не містять усього набору характеристик, або не містять певних ключових характеристик. Можливість побудови моделей, які можуть працювати з відсутніми характеристиками, буде досліджена у цій роботі, проте результати кажуть про те, що фільтрація є найбільш оптимальним вирішенням цієї проблеми. Особливо, якщо більшість об'яв структуровано правильно.

У даній роботі третій недолік вирішується повним ігноруванням будь-яких суб'єктивних характеристик. Це, безумовно, може зменшити кількість інформації щодо об'єкта, проте нівелює суб'єктивність таких оцінок.

Саме останньому недоліку з цього списку буде приділено особливу увагу у цій роботі. Спочатку буде проаналізовано вплив таких помилок на роботу і оцінку ефективності алгоритмів з оцінювання вартості нерухомості. Після цього будуть запропоновані деякі підходи для фільтрації навчальної і тестової вибірки, також буде проаналізована їхня ефективність. Тут варто відмітити, що це проблема не лише фільтрації об'яв з помилками, це більш глобальна проблема, яка може бути присутня і в даних реальних угодах купівлі-продажу нерухомості.

Деякі з об'єктів нерухомості мають ексклюзивний характер, наприклад, величезна двоповерхова студія з дизайнерським ремонтом, яка були отримана в результаті об'єднання кількох квартир у старому домі в не найкращому районі міста, буде коштувати значно дорожче порівняно з сусідніми об'єктами.

Проте його наявність у навчальній вибірці разом з кількома іншими подібними об'єктами може сильно вплинути на оцінку вартості усіх об'єктів у цьому районі, що призведе до збільшення похибки в роботі алгоритму при його тестуванні. А наявність таких даних у тестовій вибірці призводить до того, що алгоритм, який демонструє високі результати при оцінюванні інших об'єктів, отримає велику похибку при тестуванні всієї вибірки, що призведе до хибного висновку про загальну неефективність алгоритму.

Мета дослідження

Головною метою даної роботи є знаходження ефективного підходу для фільтрації даних з відкритих джерел з ринку нерухомості, який би дозволив отримати набори даних, на яких можливо ефективно навчати системи для оцінки вартості нерухомості.

Результати дослідження

Для перевірки ефективності роботи з отриманими вибірками даних буде використовуватися два алгоритми: k-середніх та штучні нейронні мережі. Метод k-середніх часто застосовується як базовий алгоритм для порівняння різних алгоритмів для оцінювання вартості нерухомості [5]. Штучні нейронні мережі також в останній час стали часто використовуватися для цих задач [6].

Оскільки в Україні на ринку нерухомості переважно представлені квартири у багатоквартирних будинках, тому для оцінювання краще використовувати не власне вартість об'єкта, а вартість за квадратний метр. Це викликано тим, що у квартир, зазвичай, ціна лінійно залежить від площі (у двох квартир, розташованих в одному будинку, вартість майже пропорційна до їхньої площі), а відповідний коефіцієнт (тобто вартість за квадратний метр) формується на основі характеристики квартир. Для новобудов, як правило, вказують лише вартість за квадратний метр, вартість конкретних об'єктів у будинках формується на його основі. А, наприклад, для будинків з земельною ділянкою оцінювання вартості є більш складним процесом, оскільки важливим фактором є і площа будинку, і площа ділянки (вплив характеристик об'єкта є також більш складним, ніж для квартир. Тому, тут і надалі, під вартістю об'єкта розуміється вартість за квадратний метр, якщо не сказано іншого.

Для методу k-середніх використовувалися лише координати об'єктів. За його допомогою усі існуючі об'єкти були розділені на 400 кластерів. При оцінюванні нового об'єкту знаходився кластер, до якого він належить, а його підсумкова оцінка визначалась як середня вартість усіх об'єктів у відповідному кластері.

Штучна нейронна мережа, що використовується для тестування, – це багатошаровий персептрон Розенблата, що складається з вхідного (9 нейронів), одного прихованого (100 нейронів) і вихідного шару (1 нейрон).

Усі характеристики об'єкта перетворювалися на вектор з 9 координат, кожній з яких відповідав нейрон вхідного шару, значення, яке передавалося на нейрон, було значенням відповідної характеристики.

Прихований шар містив 100 нейронів, кожен з яких мав 10 зв'язків: з усіма нейронами вхідного шару, а також додатковий постійний вхід. Функцією активації для нейронів цього рівня є сигмоїдальна функція. Також варто зазначити, що після певної кількості нейронів подальше збільшення їхньої кількості на прихованому шарі перестає покращувати результати ШНМ.

Вихідний шар складався з 1 нейрона, сигналом якого і є загальний результат. Цей нейрон має 101 зв'язок: з усіма нейронами прихованого шару і додатковий

постійний вхід. Оскільки результат неможливо обмежити в загальному випадку, тому для функції активації використовується лінійна функція.

Для зворотного розповсюдження помилки використовувався алгоритм Левенберга-Маркарда. Функцією оцінки роботи мережі було обрано середньоквадратичну помилку. Нейронна мережа навчалася до тих пір, поки 6 ітерацій поспіль не давали покращення оцінки.

Для перевірки ефективності роботи алгоритмів використовувалися наступні метрики:

1. Середня абсолютна помилка:

$$\frac{\sum|f(x_i)-p(x_i)|}{n},$$

де $f(x_i)$ – результат оцінювання відповідного алгоритму, $p(x_i)$ – реальна вартість об’єкта, n – кількість об’єктів у вибірці.

2. Середня абсолютна помилка у відсотках:

$$100\cdot\frac{\sum\frac{|f(x_i)-p(x_i)|}{p(x_i)}}{n}.$$

3. Середня квадратична помилка:

$$\sqrt{\frac{\sum\left(f(x_i)-p(x_i)\right)^2}{n}}.$$

4. Середня квадратична помилка у відсотках:

$$\sqrt{\frac{\sum\left(\frac{f(x_i)-p(x_i)}{p(x_i)}\right)^2}{n}}.$$

При роботі з фільтрацією помилкових даних будуть використовуватися лише об’єкти, що мають усі характеристики. Після цього буде порівняно роботу алгоритмів з об’єктами, що не мають певних характеристик.

Як база для отримання вибірок даних використовується один і той самий набір, який було отримано в результаті збору інформації протягом тижня з відкритих джерел з об’явами про продаж нерухомості.

Дані неперервно збиралися протягом тижня (до того моменту, коли сумарний розмір вибірки склав рівно 40000). Кожні 15 хвилин усі об’яви з сайту оброблялися і нові дані додавалися до вибірки, що наповнювалась. Перелік характеристик об’єктів і деталі їхнього статистичного розподілу наведено у таблиці 1. Після цього були відфільтровані усі об’єкти, що не містили загальної вартості або площі. Якщо у об’єкта була відсутня певна характеристика, то її значення прирівнювалось до нуля.

Тут варто звернути увагу на декілька моментів. По-перше, ми бачимо, що деякі об’єкти є гарантовано невалідними (максимальна площа житлової частини більша, ніж максимальна загальна площа, максимальна площа кухні на рівні максимальної загальної площі). По-друге, мінімальна вартість занадто мала (найімовірніше, це пояснюється тим, що деякі об’яви з оренди квартири опинилися не в тому розділі). По-третє, максимальна вартість об’єктів так само не виглядає реалістичною (це або помилкові дані, або якісь ексклюзивні варіанти, які неможливо оцінювати загальним підходом).

Незважаючи на це середні і медіанні значення досить реалістичні, стандартне відхилення має велике значення лише для загальної вартості, що каже про

можливість ефективно використовувати цю вибірку (після деякої фільтрації) для роботи алгоритмів оцінки вартості нерухомості.

Таблиця 1. Статистичний розподіл даних без фільтрації

Характеристика	Мінімум	Максимум	Середнє	Медіана	Стандартне відхилення
Загальна площа	1	825	59.2785	52	39.1379
Кухні	1	700	10.644	9	10.1217
Житлова частина	1	985	34.3533	30	26.7785
Кількість кімнат	1	123	2.1029	2	1.9255
Широта	36.4996	59.9164	48.8653	48.9329	1.6211
Довгота	22.2	88.5508	31.8675	30.7450	3.4654
Поверх	1	59	4.9993	4	4.1718
Загальна кількість поверхів	1	70	8.9793	9	5.8639
Загальна вартість	40	12500000	42488.5665	27000	134404.5475
Вартість за кв. м	1.1207	215517.24	672.7524	550	1493.4591

Процес навчання і тестування для обох алгоритмів відбувався наступним чином: відповідна вибірка ділилася на дві зі співвідношенням кількості елементів 1:9. Більша частина використовувалася для навчання, менша для тестування.

Застосувавши обидва алгоритми, на базовій вибірці були отримані результати, що відображені у таблиці 2. Тут і надалі ME – середня помилка, MP – середня помилка у відсотках, SE – середня квадратична помилка, SP – середня квадратична помилка у відсотках.

Таблиця 2. Результати роботи алгоритмів на вибірці без фільтрації

Алгоритм	ME	SE	MP	SP
ШНМ	265.8527	538.3815	228.6466	2447.4502
k-середніх	301.7086	419.376	231.5515	1552.395

Як видно, результати були досить невтішними. Середні помилки складають майже половину медіани, а середні квадратичні помилки мають великі значення. Тут варто звернути увагу на велике значення помилки у відсотках. Це викликано тим, що без фільтрації у вибірку потрапляє певна кількість об'яв з оренди з малою вартістю. Оцінка, яку повертають алгоритми, - це вартість продажу, яка в багато разів перевищує вартість оренди, при цьому середня абсолютна помилка є значною, проте вона має порядок середньої вартості (~600), у той час, коли помилка у відсотках може складати ~2000-3000. Саме тому важливо розглядати обидві ці метрики.

Найбільш очевидним підходом до фільтрації є встановлення меж вартості об'єктів. Як базовий інтервал було визначено інтервал від 80 до 2000. У таблиці 3 наведено статистичний розподіл цієї вибірки. Як видно, розподіл за усіма характеристиками (крім вартості) майже не змінився, при цьому стандартне відхилення вартості значно зменшилося (тобто розподіл став більш рівномірним). Крім того, варто зазначити, що розмір вибірки зменшився з 37385 елементів до 35633 елементів.

Окрім цього підходу були використані наступні:

1. Кожен з алгоритмів навчився на нефільтрованих даних і відкинув 5% з них з

найбільшою помилкою. Для кожного з алгоритмів була створена своя вибірка.

2. Вибірка була відфільтрована за вартістю в інтервалі між персентилями (2.5 і 97.5) вартості.
3. Для ШНМ була створена вибірка, в якій визначені всі характеристики (у деяких об'єктах базової вибірки могли бути відсутні одна або декілька характеристик).
4. Для ШНМ була створена вибірка, в якій відсутні характеристики було замінено на середні значення цих характеристик серед усіх об'єктів.

Таблиця 3. Статистичний розподіл даних з фільтрацією за вартістю за квадратний метр

Характеристика	Мінімум	Максимум	Середнє	Медіана	Стандартне відхилення
Загальна площа	9	825	56.8525	52	27.435
Кухні	1	455	10.3673	9	8.3697
Житлова частина	1	985	33.4385	30	22.6877
Кількість кімнат	1	79	2.0707	2	1.1594
Широта	44.4071	59.9164	48.8524	48.9236	1.6237
Довгота	22.2	88.5508	31.8277	30.745	3.4652
Поверх	1	59	5.01897	4	4.1812
Загальна кількість поверхів	1	70	8.9887	9	5.8443
Загальна вартість	1840	950000	37637.6717	28000	35122.7594
Вартість за кв. м	80.3571	1997,9167	631,3745	553,6852	332,6722

Результати тестування відповідних вибірок наведено у таблицях 4 і 5. Номери вибірок наступні:

1. Базова вибірка без фільтрації.
2. Вибірка, відфільтрована за вартістю (від 80 до 2000).
3. Вибірка, відфільтрована за вартістю (від персентиля 2.5 до персентиля 97.5).
4. Вибірка, відфільтрована k-середніми (5% найбільших помилок).
5. Вибірка, відфільтрована ШНМ (5% найбільших помилок).
6. Вибірка, відфільтрована за вартістю (від 80 до 2000) з об'єктами з усіма характеристиками.
7. Вибірка, відфільтрована за вартістю (від 80 до 2000), у якій усі відсутні значення характеристик замінені на середні для відповідної характеристики.

Таблиця 4. Результати роботи ШНМ для різних вибірок

Вибірка	ME	SE	MP	SP
1	265.8527	538.3815	228.6466	2447.4502
2	168.0008	240.6054	33.5472	55.1093
3	173.9537	261.7203	37.9937	74.3714
4	162.8477	227.4383	155.7282	972.4003
5	202.5184	311.9653	107.5759	707.7789
6	169.4322	258.7070	34.5087	98.7667
7	166.6703	247.5252	32.6885	53.9494

Таблиця 5. Результати роботи методу k-середніх для різних вибірок

Вибірка	ME	SE	MP	SP
1	301.7086	419.376	231.5515	1552.395
2	246.2272	342.7200	49.4481	82.0672
3	269.8851	331.9069	70.7983	125.5351
4	264.2601	314.688	239.1907	1299.1948
5	617.8787	740.9010	93.7950	98.2811

Як видно з таблиць 4 і 5, більшість із запропонованих підходів до фільтрації значно покращує результати роботи обох алгоритмів. Найкращим із запропонованих виявився фільтр за вартістю у діапазоні від 80 до 2000 (2), при цьому він навіть обійшов майже аналогічний фільтр за персентилями (3).

Цікаво, що фільтр за допомогою методу k-середніх виявився кращим для обох алгоритмів, аніж аналогічний фільтр за допомогою ШНМ.

Велике значення помилки у відсотках для деяких фільтрів (4 і 5) можна пояснити тим фактом, що вони не відсіювали об’єкти з малим значенням вартості (оскільки вони не мають значної помилки). А ось фільтри, що відсіювали подібні об’єкти (2 і 3), показали ефективні результати і для звичайної помилки, і для помилки у відсотках.

Якщо ж розглядати додаткові вибірки для ШНМ, то можна помітити, що більш строга фільтрація (наявність усіх характеристик у об’єктів) не призводить до покращення роботи алгоритму, значення у 2 і 6 майже не відрізняються і навіть 2 перевершує 6 за всіма параметрами. При цьому зміна нулів на середні значення характеристик незначно покращила три з чотирьох метрик. Проте покращення досить незначне, що не дає можливості говорити про однозначну ефективність запропонованого підходу.

Висновки

Проблема побудови актуальних і якісних вибірок навчальних і тестових даних є дуже актуальною при роботі над оцінюванням вартості об’єктів нерухомості. Як було показано у роботі, дані, що збираються з оголошень з відкритих джерел, не можна використовувати напряму для роботи відповідних алгоритмів.

Запропоновані в роботі підходи для фільтрації наборів даних показали себе достатньо ефективно: результати роботи базових алгоритмів значно покращилися на відфільтрованих вибірках, що дозволяє говорити про можливість їхнього використання на практиці і для подальшої розробки більш ефективних алгоритмів оцінювання вартості нерухомості.

Частина роботи щодо оцінювання об’єктів з відсутніми характеристиками для ШНМ показала, що наявність в об’єктів усіх характеристик не впливає на якість оцінювання, проте результати роботи запропонованої оптимізації не дають можливості однозначно говорити про її ефективність.

Література

1. Helbich M. et al. Data-driven regionalization of housing markets //Annals of the Association of American Geographers. – 2013. – Т. 103. – №. 4. – S. 871-889.
2. Zurada J., Levitan A., Guan J. A comparison of regression and artificial intelligence methods in a mass appraisal context //Journal of Real Estate Research. -- 2011.
3. Baranzini A., Schaerer C. A sight for sore eyes: Assessing the value of view and land use in the housing market //Journal of Housing Economics. -- 2011. -- Т. 20. -- №. 3. - S. 191-199.

4. Kuşan H., Aytekin O., Özdemir İ. The use of fuzzy logic in predicting house selling price //Expert systems with Applications. -- 2010. -- T. 37. -- №. 3. -- S. 1808-1813.
5. Bourassa S., Cantoni E., Hoesli M. Predicting house prices with spatial dependence: A comparison of alternative methods //Journal of Real Estate Research. – 2010.
6. Kauko T. A comparative perspective on urban spatial housing market structure: Some more evidence of local sub-markets based on a neural network classification of Amsterdam //Urban Studies. – 2004. – T. 41. – №. 13. – S. 2555-2579

RESUME

A.O. Shcherba

Dataset optimization for real estate valuation

Being very practical task real estate valuation requires a lot of real market data. There is no open central database with such data in Ukraine, so the only effective source for real estate prices are advertisement sites, which also contain invalid and erroneous data.

This article analyzes the main problems of such datasets (price is often not final, not all information is present, mistakes in characteristics, e.g. wrong format, wrong values, etc.) and their influence on learning and testing.

The main purpose of this article is to find the effective ways of filtering such data, which makes possible to create large datasets for real estate valuation algorithms.

Several ways of filtering were proposed: filter with fixed range, filter with percentile range, filtering of the worst results of the specific algorithm. The best results were achieved by the first two approaches. The main problem of latter ones was their performance on elements with small prices. Absolute error for such elements is often small, but error in percent can be much larger.

Also, three ways for handling missing information (some characteristics are not set for some elements in dataset) for artificial neural networks are analyzed: filtering such elements, setting value of such characteristics to 0 and to mean value of that characteristic. The testing results have not shown significant difference in neural network performance on these datasets.

The results of both algorithm (neural network and k-means) with range filtering of initial dataset are good enough, which opens the prospects for researching new more efficient algorithms for real estate valuation on these datasets and the new ones, created with such filtering.

Надійшла до редакції 06.10.2017